EXPRESS MAIL

CERTIFICATE OF MAILING

Patent Application for:

Applicants: Richard V. Cox
Hong Kook Kim

Atty. No: 1999-0767-1

Title: Frame Erasure Concealment Technique for a Bitstream-Based Feature Extractor

I hereby certify that this correspondence is being deposited with the United States Postal Service with sufficient postage as EXPRESS MAIL in an envelope addressed to

Commissioner for Patents

Washington, D.C. 20231

On _____ 12/5/00 _____

EXPRESS MAIL TRACKING NO. EF172637792 US

Type or Print name of person signing this certificate: Wendy W. Koba, Esq.

Signature: Wendy W. Koba

FRAME ERASURE CONCEALMENT TECHNIQUE FOR
A BITSTREAM-BASED FEATURE EXTRACTOR

Cross-Reference to Related Applications

5   This application claims the priority of Provisional Application No. 60/170,170, filed December 10, 1999.

Technical Field

The present invention relates to automatic speech recognition and, more
10   particularly, to a frame erasure concealment technique for use with a bitstream-based feature extraction process in wireless communication applications.

Background of the Invention

In the provisioning of many new and existing communication services, voice
15   prompts are used to aid the speaker in navigating through the service. In particular, a speech recognizing element is used to guide the dialogue with the user through voice prompts, usually questions aimed at defining which information the user requires. An automatic speech recognizer is used to recognize what is being said and the information is used to control the behavior of the service rendered to the user.

20   Modern speech recognizers make use of phoneme-based recognition, which relies on phone-based sub-word models to perform speaker-independent recognition over the telephone. In the recognition process, speech "features" are computed for each incoming frame. Modern speech recognizers also have a feature called "rejection". When rejection exists, the recognizer has the ability to indicate that what was uttered does not correspond
25   to any of the words in the lexicon.

The users of wireless communication services expect to have access to all of the services available to the users of land-based wireline systems, and to receive a similar quality of service. The voice-activated services are particularly important to the wireless subscribers since the dial pad is generally away from sight when the subscriber listens to
30   a vocal prompt, or is out of sight when driving a car. With speech recognition, there are virtually no restrictions on mobility, because callers do not have to take their eyes off the road to punch in the keys on the terminal.

Currently, one area of research is focusing on the front-end design for a wireless speech recognition system. In general, many prior art front-end designs fall into one of two categories, as illustrated in FIG. 1. FIG. 1(a) illustrates an arrangement 10 including a speech encoder 12 at the transmitting end, a communication channel 14 (such as a

5   wireless channel) and a speech decoder 16 at the receiving end. The decoded speech is thereafter sent to *EAR* and also applied as an input to a speech recognition feature extractor 18, where the output from extractor 18 is thereafter applied as an input to an automatic speech recognizer (not shown). In a second arrangement 20 illustrated in FIG. 1(b), a speech recognition feature encoder 22 is used at the transmitting end to allow for

10   the features themselves to be encoded and transmitted over the (wireless) channel 24. The encoded features are then applied as parallel inputs to both a speech decoder 26 and a speech recognition feature extractor 28 at the receiving end, the output from feature extractor 28 thereafter applied as an input to an automatic speech recognizer (not shown). This scheme is particularly useful in Internet access applications. For example, when the

15   mel-frequency cepstral coefficients are compressed at a rate of approximately 4kbit/s, the automatic speech recognizer (ASR) at the decoder side of the coder exhibits a performance comparable to a conventional wireline ASR system. However, this scheme is not able to generate synthesized speech of the quality produced by the system as shown in FIG. 1(a).

20   In speech coding, channel impairments are modeled by bit error insertion and frame erasure insertion devices, where the number of bit errors and frame erasures depends primarily on the noise, co-channel and adjacent channel interference, as well as frequency-selective fading. Fortunately, most speech coders are combined with a channel coder, where a "frame erasure" is declared if any of the most sensitive bits with

25   respect to the channel is in error. The speech coding parameters of an erased frame must then be extrapolated in order to generate the speech signal for the erased frame. A family of error concealment techniques are known in the prior art and can generally be defined as either "substitution" or "extrapolation" techniques. In general, the parameters of the erased frames are reconstructed by repeating the parameters of the previous frame with

30   scaled-down gain values. In conventional speech recognition systems, a decoded speech-

based front-end uses the synthesized speech for extracting a feature. However, in a bitstream-based front-end, the parameters themselves are present.

The need remaining in the prior art, therefore, is to provide a technique for handling frame erasures in a bitstream-based front end speech recognition systems.

5

Summary of the Invention

The need remaining in the prior art is addressed by the present invention, which relates to automatic speech recognition and, more particularly, to a frame erasure concealment technique for use with a bitstream-based feature extraction process in 10 wireless communication applications.

In accordance with the present invention, an error in a frame is declared if the Euclidean distance between the line spectrum pair (LSP) coefficients in adjacent frames is less than or equal to a predefined threshold T. In such a case, one of the frames in then simply deleted from the bitstream. In particular, and based on the missing feature theory, 15 a decoding algorithm is reformulated for the hidden Markov model (HMM) when a frame erasure is detected.

Other and further features and advantages of the present invention will become apparent during the course of the following discussion and by reference to the accompanying drawings.

20

Brief Description of the Drawings

Referring now to the drawings,

FIGs. 1(a) and (b) illustrate, in simplified block diagram form, two prior arrangements for exemplary wireless automatic speech recognition systems;

25 FIG. 2 illustrates, in block diagram form, the components utilized in a speech recognition system of the present invention;

FIG. 3 is a simplified flow chart illustrating the feature extraction process associated with the IS-641 speech coder;

FIG. 4 contains a diagram of the procedure for extracting feature parameters 30 directly from the bitstream in accordance with the present invention;

3

FIG. 5 illustrates an exemplary arrangement for modeling the efficacy of the proposed feature extractor of the present invention when compared with prior art arrangements;

FIG. 6 illustrates a process of the present invention used to obtain additional 5 "voiced" and "unvoiced" information from the bitstream;

FIG. 7 contains exemplary trajectories of adaptive codebook gain (ACG) – voiced, and fixed codebook gain (FCG) – unvoiced – parameters for speech after processing by an IS-641 speech coder;

FIG. 8 illustrates various speech waveforms associated with the implementation 10 of an exemplary speech enhancement algorithm in association with the feature extraction process of the present invention;

FIG. 9 contains graphs illustrating the word error rate (WER) associated with various frame erasure techniques; and

FIG. 10 illustrates the ratios of processing time between a conventional 15 extrapolation frame erasure technique and the frame deletion method of the present invention.

Detailed Description

20   A bitstream-based approach for providing speech recognition in a wireless communication system in accordance with the present invention is illustrated in FIG. 2. As shown, a system 30 utilizes a conventional speech encoder 32 at the transmission end, where for explanatory purposes it will be presumed that an IS-641 speech coder is used, 25 however, various other coders also function reliably in the arrangement of the present invention (in particular, code-excited linear prediction – CELP encoders). The encoded speech thereafter propagates along a (wireless) communication channel 34 and is applied as simultaneous inputs to both a speech decoder 36 and a speech recognition feature extractor 38, where the interaction of these various components will be discussed in detail 30 below.

FIG. 3 includes a simplified block diagram of the linear predictive coding (LPC) analysis associated with speech coding performed using an IS-641 speech coder. As shown, the speech coder first removes undesired low frequency components from the speech signal by application of a highpass filter 40 with a cutoff frequency of, for example, 80 Hz. The filtered speech is then applied as an input to an autocorrelation function using an asymmetric window 42, where one side of the window is half of a Hamming window and the other half is a quarter period of the cosine function. The particular shape realized by this asymmetric window is due to the limited lookahead of the speech coder for minimizing the delay for real applications. Subsequent to the windowing, two additional processes 44 are applied to the autocorrelated signal. One is defined as lag-windowing and the other is white noise correction. The former helps to smooth the LPC spectrum so as to exhibit no sharp peaks. The white noise correction provides the effect of adding noise to the speech signal and thus avoids modeling the anti-aliasing filter response at high frequencies with the LPC coefficients. Finally, a conventional LPC recursion is performed (block 46) with the modified autocorrelation sequence output from block 44 to form the line spectrum pair (LSP) coefficient output. A speech encoder 48 then quantizes the LSP coefficients and transmits them as the "bit stream" output to a decoder (not shown). When the LSP coefficients are recovered at the decoder side, the decoded LSP's will be somewhat different from the unquantized LSP's, depending on the performance of the spectral quantizer within speech encoder 48.

With this understanding of the encoding process within an IS-641 speech encoder, it is possible to study in detail the bitstream recognition process of the present invention. Referring to FIG. 4, a procedure is illustrated for extracting cepstral coefficients from the bitstream of an IS-641 speech coder (the bitstream being, for example, the output of the arrangement illustrated in FIG. 3). A single frame is illustrated in FIG. 4 and contains two major divisions. As shown, bits 1-26 are used for the LSP quantization while the remaining bits 27 – 148 are used for all residual information. In the process, the decoded LSP's are decoded from the first 26 bits using a inverse quantizer 50 where, for example, these LSP's represent the spectral envelope of a 30ms speech segment with a frame rate of 50 Hz. In order to match to the frame rate with that of a conventional speech recognition front-end, the output from inverse quantizer 50 is interpolated with the LSP's

of the previous frame (block 52) to convert the frame rate to 100 Hz. Next, cepstral coefficients of order 12 are obtained by performing an LSP to LPC conversion, followed by an LPC to CEP conversion (block 54). By applying a bandpass filter 56 to the cepstral coefficients, a set of twelve weighted coefficients is obtained. The residual signal from

5   bits 27 – 148, identified as "pitch information" (bits 27 – 52), "algebraic codebook information" (bits 53 – 120) and "codebook gains" (bits 121 – 148), are also decoded. An energy parameters is then computed by taking the logarithm to the square-sum of the residual (20 ms).

      Although this description is particular to the IS-641 speech coder, it is to be

10   understood that the feature extraction process of the present invention is suitable for use with any code-excited linear prediction (CELP) speech coder.

      The model illustrated in FIG. 5 can be used to measure the efficacy of the bitstream-based system of the present invention with various other ASR techniques. Illustrated in general is an IS-641 speech encoder 60, including an analysis module 62

15   and a quantizer 64. An IS-641 speech decoder 66 is also shown, separated from IS-641 speech encoder by an ideal channel 68. Included within speech decoder 66 is an inverse quantizer 68 and a synthesis module 70. A conventional speech signal is applied as an input to analysis module 62 and the decoded speech will ultimately exit from synthesis module 70. The location of reference point C0 corresponds to the placement of a

20   conventional wireline speech recognition system. At reference point C1, ASR is performed on a speech signal coded using IS-641 (corresponding to the arrangement shown in FIG. 1(a)). In order to evaluate the ASR performance of the LPC analysis method (associated with FIG. 1(b)), an ASR at location C2 can be used with the unquantized LSP's as generated by LPC recursion process (block 46 of FIG. 3). Lastly,

25   an ASR positioned at location C3 (directly converting the bitstream output of the IS-641 coder into the speech recognition feature set) can then be used to analyze the bitstream-based front end arrangement of the present invention.

      Tables I and II below include the speech recognition accuracies for each ASR pair, where "Cx/Cy" is defined as an ASR that is trained in Cx and then tested in Cy:

30

Feature	Word Accuracy (%)	Word Error (%)			String Accuracy (%)
		Sub.	Del.	Ins.	
C0/C0 (10ms)	96.17	1.46	0.78	1.59	68.48
C0/C0 (20 ms)	95.81	1.60	0.76	1.83	66.06
C0/C1	95.16	2.09	0.95	1.79	62.31
C1/C1	94.75	2.38	1.01	1.86	60.20

Table I

Feature	Word Accuracy (%)	Word Error (%)			String Accuracy (%)
		Sub.	Del.	Ins.	
C2/C2	96.23	1.43	0.71	1.63	68.91
C3/C3	95.81	1.68	0.82	1.69	66.48

Table II

Table I includes a comparison of the recognition accuracy for each of the conventional front-ends, using the ASR location identifiers described above in association with FIG. 5. Alternatively, Table II provides a listing of the recognition accuracy of bitstream-based front-end speech recognition performed in accordance with the present invention as located in either the encoder side or decoder side of the speech coder arrangement. Referring to Table II, comparing the C2/C2 results with the C3/C3 results, it is shown that the word and string accuracies of C3/C3 are decreased by 12% and 8%, respectively (results comparable to C0/C0 with linear interpolation). It has been determined that this degradation is caused mainly by the LSP quantization in the IS-641 speech coder. Therefore, the arrangement of the present invention further requires a method of compensating for the LSP quantization effects. In accordance with the present invention, unvoiced/voiced information is incorporated in the feature set so that the feature set as a whole can compensate for the quantization effect.

As mentioned above, in addition to the spectral envelope, a speech coder models the excitation signal as the indices and gains of the adaptive and fixed codebooks, where these two gains represent the "voiced" (adaptive codebook gain – ACG) and "unvoiced"

(fixed codebook gain – FCG) information. These parameters are quantized and then transmitted to the decoder. Therefore, in accordance with the present invention, it is possible to obtain the voiced/unvoiced information directly from the bitstream. FIG. 6 illustrates an exemplary process of extracting these additional "voiced" and "unvoiced"

5 parameters in the bitstream-based front-end of the invention. As shown, bits 121 – 148 in an exemplary frame (the "gain" information as shown in FIG. 4) are further divided into four subframes, denoted SF0, SF1, SF2, and SF3, where the ACG (voiced) and FCG (unvoiced) values are computed for each subframe. Therefore, four ACG values and four FCG values are determined for each frame (blocks 70, 72). In order to generate speech

10 recognition feature parameters from these gains, the following equations are used:

$$ACG(i) = \sum_{k=0}^{1} g_p^2(2i+k), i = 0,1 \qquad (1)$$

$$FCG(i) = \gamma 10 \log_{10}\{\sum_{k=0}^{1} g_c^2(2i+k)\}, i = 0,1 \qquad (2)$$

where $g_p(i)$ and $g_c(i)$ are defined as the ACG and FCG values of the i-th subframe. In order to add the ACG and FCG values into the feature vector and maintain the same vector dimension as before, two of the twelve LPC cepstra values in the baseline are

15 eliminated.

FIG. 7 illustrates an example of the trajectories of the adaptive codebook gain and fixed codebook gain for a speech waveform after it has been processed by an IS-641 speech coder. FIG. 7(a) is an illustration of an exemplary digit string, and FIG. 7(b) is the normalized energy parameter associated with this digit string. FIGs. 7(c) and 7(d)

20 illustrated the ACG and FCG parameters, respectively, for this string. As can be seen, both the ACG and FCG exhibit temporal fluctuations. These fluctuations can be reduced by applying a smoothing technique (such as median filtering, illustrating as blocks 74 and 76 in FIG. 6). As with the typical energy parameters in speech coding, a weighting function (denoted as block 78 in FIG. 6 and defined as γ in Eq. (2)) can be added to the

25 filtered FCG parameters, where the weighting function is chosen to control the effect of the FCG parameters relative to the various other parameters. In one exemplary arrangement, γ may be equal to 0.1.

Table III, included below, illustrates the improved results from incorporating the ACG and FCG parameters into the feature set. Compared with the baseline, the new feature set reduces the word and string error rates by 10% for each. Referring back to Tables I and II, these results for the arrangement of the present technique of incorporating

5 ACG and FCG are now comparable to the conventional prior art models.

Feature	Word Accuracy (%)	Word Error (%)			String Accuracy (%)
		Sub.	Del.	Ins	
C3: Wireless Baseline	95.81	1.68	0.82	1.69	66.48
C3-1: LPC-CEP, AFG, FCG	95.96	1.84	0.80	1.39	67.84
C3-2: Median Smoothing	95.98	1.86	0.78	1.38	68.69
C3-3: Gain Scaling	96.24	1.69	0.72	1.35	69.77

Table III

In order to properly analyze these recognition results, it is possible to use

10 hypothesis tests for analyzing word accuracy (using matched-pair testing) and string accuracy (using, for example, McNemar's testing). A complete description of McNemar's testing as used in speech recognition can be found in the article entitled *"Some statistical issues in the comparison of speech recognition algorithms"*, by L. Gillick and S. Cox appearing in <u>Proceedings of the ICASSP</u>, p. 532 et seq., May 1989.

15 For matched-pair testing, the basic premise is to test whether the performance of a system is comparable to another or not. In other words, a hypothesis H_0 is constructed as follows:

$$H_0 : \mu_A - \mu_B = 0, \qquad (3)$$

where μ_A and μ_B represent the mean values of the recognition rates for systems A and B,

20 respectively. Alternatively, to test the string accuracy, McNemar's test can be used to test the statistical significance between the two systems. In particular, the following

"null" hypothesis is tested: *If a string error occurs from one of the two systems, then it is equally likely to be either one of the two.* To test this, N_{01} is defined as the number of strings that system A recognizes correctly and system B recognizes incorrectly. Similarly, the term N_{10} will define the number of strings that system A recognizes

5   incorrectly and system B recognizes correctly. Then, the test for McNamara's hypothesis is defined by:

$$W = \frac{|N_{10} - k/2| - 1/2}{\sqrt{k/4}} \qquad (4)$$

where $k = N_{01} + N_{10}$.

    As an example, these test statistics can be computed for a "wireless baseline"

10   system (C3) and bitstream-based front-end system (C3-3) of the present invention, including both ACG and FCG, using the data from Table III.

Features		Matched-pairs	McNamara
A	B	(W)	(W)
C3-3	C3	1.965	2.445

Table IV

15

    The results of these computations are shown above in Table IV, where from these results it is clear that the incorporation of ACG and FCG in the arrangement of the present invention provides significantly improved recognition performance over the baseline with a confidence of 95%. Moreover, Table V (shown below) illustrates that the proposed

20   front-end of the present invention yields comparable word and string accuracies to conventional wireline performance.

Features		Matched-pairs	McNamara
A	B	(W)	(W)
C0	C1	3.619	3.607
C3-3	C1	3.914	4.388
C3-3	C0	0.328	0.833

Table V

5   The performance of the bitstream-based front end of a speech recognizer can also be analyzed for a "noisy" environment, such as a car, since oftentimes a wireless phone is used in such noisy conditions. To simulate a noisy environment, a car noise signal can be added to every test digit string. That is, the speech recognition system is trained with "clean" speech signals, then tested with noisy signals. The amount of additive noise can

10   be measured by the segmental signal-to-noise ratio (SNR). Table VI, below, shows the recognition performance comparison when the input SNR varies from 0 dB to 30 dB in steps of 10 dB.

SNR (db)		0	10	20	30	∞
C0/	Word	14.30	61.82	85.84	95.73	96.17
C0	String	0.0	0.51	23.07	65.49	68.48
C0/	Word	21.18	65.59	85.47	94.29	95.16
C1	String	0.0	0.51	19.96	55.75	62.32
C3-3	Word	16.82	67.28	90.64	95.28	96.24
C3-3	String	0.0	3.62	41.59	63.79	69.77

15   *Table VI*

As shown, for an SNR above 20dB, the bitstream-based front-end arrangement of the present invention (C-3/C-3) shows a better performance than the conventional wireless front end. However, its performance is slightly lower than the conventional wireline front end. With lower values of SNR, the arrangement of the present invention does not

compare as favorably, particularly due to the fact that the inventive front-end utilizes voicing information, but the speech coder itself fails to correctly capture the voicing information at low levels of SNR.

5   The utilization of a speech enhancement algorithm with the noisy speech signal prior to speech coding, however, has been found to improve the accuracy of the extracted voicing information. An exemplary speech enhancement algorithm that has been found useful with the processing of noisy speech is based on minimum mean-square error log-spectral amplitude estimation and has, in fact, been applied to some standard speech coders. FIG. 8 illustrates speech waveforms implementing such enhancement under a

10   variety of conditions. In particular, a "clean" speech waveform is shown in FIG. 8(a). FIG. 8(b) shows the waveform decoded by a conventional IS-641 speech coder. The "noisy" speech (e.g., contaminated by additive car noise), whose SNR is 20 dB, is shown in FIG. 8(c), and its decoded speech signal is displayed in FIG. 8(d). This particular type of speech enhancement is applied to the "noisy" signal waveform of FIG. 8(e), where the

15   speech coding is then performed after the enhancement, the result being shown in FIG. 8(f), which shows that the noise signal is removed by applying the speech enhancement algorithm.

As mentioned above, channel impairments can be modeled by bit error insertion and frame erasure insertion devices, where the number of bit errors and frame erasures

20   depends mainly on the noise, co-channel and adjacent channel interference, and frequency selective fading. Fortunately, most speech coders are combined with a channel coder. The most sensitive bits are thus strongly protected by the presence of the channel coder. A "frame erasure" is declared if any of the most sensitive bits with respect to the channel is in error. In the context of the bitstream-based arrangement of the present

25   invention, the bits for LSP (i.e., bits 1-26) and gain (i.e., bits 121 – 148) are defined as most sensitive to channel errors. Therefore, for the purposes of the present invention, it is sufficient to consider a "frame erasure" condition to exist if these bits are in error, since the recognition features in the bitstream-based front end are extracted from these bits.

In the prior art, the speech coding parameters of an erased frame are extrapolated

30   in order to generate the speech signal for the erased frame. The parameters of erased frames are reconstructed by repeating the parameters of the previous frame with scaled-

12

down gain values. In particular, the gain values depend on the burstiness of the frame erasure, which is modeled as a finite state machine. That is, if the n-th frame is detected as an erased frame, the IS-641 speech coder estimates the spectral parameters by using the following equation:

5
$$\omega_{n,i} = c\omega_{n-1,i} + (1-c)\omega_{dc,i}, \quad i = 1, ..., p$$
(5)

where $\omega_{n,i}$ is the i-th LSP of the n-th frame and $\omega_{dc,i}$ is the empirical mean value of the i-th LSP over a training database and c is a forgetting factor set to a value of 0.9. The ACG and FCG values are obtained by multiplying the predefined attenuation factors to the gains of the previous frame, and the pitch value is set to the same pitch value of the

10 previous frame. The speech signal, using this "extrapolation method" is then reconstructed from these extrapolated parameters.

As an alternative, the present invention proposes a "deletion method" for overcoming frame erasures in a bitstream-based speech recognition front end. Based on the missing feature theory, a decoding algorithm is reformulated for the hidden Markov

15 model (HMM) when a frame erasure is detected. That is, for a given HMM $\lambda = (A, B, \pi)$, the probability of the observation sequence $\mathbf{O} = \{o_1, ..., o_N\}$ is given by:

$$P(\mathbf{O} \mid \lambda) = \sum_{q_1,...,q_N} \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2)...a_{q_N-1 q_N} b_{q_N}(o_N)$$
(6)

where N is the number of observation vectors in $\mathbf{O}$, $(q_1, ..., q_N)$ is defined as a state sequence, and π_q is the initial state distribution. Also, the observation probability of o_n at

20 state i is represented as follows:

$$b_i(o_n) = \sum_{k=1}^{M} c_{ik} N(o_n; \mu_{i,k}; \Sigma_{ik})$$
(7)

where

$$N(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\mu/2} |\Sigma|^{1/2}} \exp\{-1/2(x-\mu)^2 \Sigma^{-1}(x-\mu)\},$$

M is the number of Gaussian mixtures, and c_{ik} is the k-th mixture weight of the i-th state.

25 The variables μ and σ define the mean vector and covariance matrix, respectively.

13

To understand the "deletion" method of frame erasure method of the present invention, presume that the l-th frame is detected as a missing frame. The first step in the deletion method is to compute the probability of only the correct observation vector sequence for the model λ. The observation vector sequence can be divided into two groups as follows:

$$\mathbf{O} = (\mathbf{O}^0, \mathbf{O}^m), \tag{8}$$

where $o_l \in \mathbf{O}^m$. From the missing feature theory, the probability to be computed can be expressed as follows:

$$P(\mathbf{O} \mid \lambda) = \int P(\mathbf{O}^c, \mathbf{O}^m \mid \lambda) d\mathbf{O}^m. \tag{9}$$

Also, for the missing observation vector o_l, it is known that:

$$\int b_i(o_l) do_l = 1. \tag{10}$$

By substituting (6) and (10) into (9), the following relationship is obtained:

$$P(\mathbf{O}^c \mid \lambda) = \sum_{q_1,\dots,q_N} \pi_{q1} b_{q1}(o_1) a_{q1q2} b_{q2}(o_2)\dots a_{q1-Nq1} a_{q1q1+n} b_{q1+1}(o_{l+\Gamma})\dots a_{qN-1qN} b_{qN}(o_N) \tag{11}$$

It is known that the transition probabilities have less effect in the Viterbi search than the observation probabilities. Therefore, it is possible to set $a_{q1-Nq1} = 1$. The above equation is then simply realized by deleting the vector o_l in the observation sequence and using the conventional HMM decoding procedure.

The deletion method of the present invention can be interpreted in terms of a VFR analysis. In particular, the Euclidean distance of the LSP's between the $(n-1)$-th and the n-th frames is given by:

$$\sum_{i=1}^{p}(\omega_{n,i}-\omega_{n-1,i})^2 = (1-c)\sum_{i=1}^{p}(\omega_{n-1,i}-\omega_{dc,i})^2 \qquad (12)$$

If the distance expressed above is less than or equal to a predefined threshold T, the two frames are assumed to be in the steady-state region and the LSP's of the n-th frame are

5   deleted in the observation sequence. Therefore, if for the threshold T the following is presumed:

$$T = (1-c)^2 \max_{[x_1,\ldots x_p]\in\Omega} \sum_{i=1}^{p}(x_i-\omega_{dc,i})^2 \qquad (13)$$

where Ω is a p-dimensional LSP vector space, all of the missing frames will be deleted.

In terms of computational complexity, it can be concluded that using the deletion

10   process of the present invention reduces the length of the observation sequence by $N(1-p_e)$, where p_e is the frame erasure rate (FER).

To simulate frame erasure conditions, error patterns depending on the FER and its burstiness can be generated for various test strings. For example, FIG. 9(a) illustrates the word error rate (WER) when the random FER varies from 3% to 20%. An FER of

15   0% is defined as a "clean" environment, where a 3% FER is considered typical of a conventional TDMA channel. At 3% FER, the WER's are increased by 6.4% and 5.3% for the bitstream-based front-end of the present invention, utilizing the conventional "extrapolation" frame erasure method and the inventive "deletion" method, respectively, where the deletion method has been found to have a higher deletion error and lower

20   insertion and substitution error than the extrapolation method.

FIG. 9(b) illustrates the WER as a function of the burstiness of the FER when the FER is 3% (the "burstiness" being defined as b for the sake of simplicity). Similar to the random FER case, the WER's of the bitstream-based front-ends are smaller than those associated with decoded speech-based front-ends. Comparing the WER performance at

25    $b=0.99$ to that under a "clean" environment, the decoded speech-based front-end increases the WER by 24.3%, while the bitstream-based front-ends with the extrapolation method and with the deletion method increase the WER by 19.7% and 22.1%, respectively. The inventive deletion method gives a slightly worse performance than the

extrapolation method when b is large since the deletion method increases the deletion errors as b increases.

FIG. 10 illustrates the ratios of processing time between the extrapolation method and the deletion method for each FER and level of burstiness. For the purposes of this graph, the processing time was calculated by performing recognition experiments overall all the test data on the same machine. As shown, the results verify that the proposed deletion method has less computational complexity than the extrapolation method.

While the exemplary embodiments of the present invention have been described above in detail, it is to be understood that such description does not limit the scope of the present invention, which may be practiced in a variety of embodiments. Indeed, it will be understood by those skilled in the art that changes in the form and details of the above description may be made therein without departing from the scope and spirit of the invention.